Optimal low-rank approximations of Bayesian linear inverse problems

L. Tenorio
Applied Mathematics and Statistics
Colorado School of Mines

September 2017
Acknowledgments

In collaboration with: Alessio Spantini (MIT) & Youssef Marzouk (MIT)

- Optimal low-rank approximations of Bayesian linear inverse problems, SISC
- Goal-oriented optimal approximations of Bayesian linear inverse problems, SISC (to appear)

Partially funded by: DOE ASCR DE-SC0003908, DE-SC0009297
Bayesian linear inverse problem
(Recall Daniela’s talk)

- $y = \text{data}$
  $x = \text{high-dimensional parameter vector}$
  $G = \text{forward operator}$

- Bayesian model: assume $\Gamma_{pr} \succ 0$ and $\Gamma_{obs} \succ 0$,
  
  $$y \mid x \sim N(Gx, \Gamma_{obs}), \quad x \sim N(0, \Gamma_{pr})$$

Posterior: $x \mid y \sim N(\mu_{pos}(y), \Gamma_{pos})$

$$\mu_{pos}(y) = \Gamma_{pos} G^\top \Gamma_{obs}^{-1} y$$

$$\Gamma_{pos} = (H + \Gamma_{pr}^{-1})^{-1} \succ 0$$

$$H = G^\top \Gamma_{obs}^{-1} G \quad \text{i.e., ‘Hessian’}$$
\( \Gamma_{\text{pos}} \) as an update of \( \Gamma_{\text{pr}} \)

- Will focus on approximations of \( \Gamma_{\text{pos}} \)
- By Gaussianity:

\[
\begin{align*}
\Gamma_{\text{pos}} & \preceq \Gamma_{\text{pr}} \\
\Gamma_{\text{pos}} &= \Gamma_{\text{pr}} - \text{Var}[\mu_{\text{pos}}(y)] \\
&= \Gamma_{\text{pr}} - KK^\top \succ 0
\end{align*}
\]

- High-dimensionality \( \Rightarrow \) low-rank of update \( KK^\top \)

This defines a natural approximation class:

\[
\mathcal{M}_r = \left\{ \Gamma = \Gamma_{\text{pr}} - KK^\top : \Gamma \succ 0, \ \text{rank}(K) \leq r \right\}
\]
Comparing approximations of $\Gamma_{\text{pos}}$

- How do we compare a chosen $\hat{\Gamma}_{\text{pos}} \in \mathcal{M}_r$ to $\Gamma_{\text{pos}}$?
  I.e., choose an appropriate loss-function $L(\hat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}})$

- Simple example: If $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, then

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \text{unbiased estimator of } \sigma^2$$

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2 = \text{M.L. estimator of } \sigma^2$$

Choose class $\mathcal{M} = \{ \hat{\sigma}_\delta^2 = \delta S^2 : \delta > 0 \}$ and loss

$$L(\hat{\sigma}^2, \sigma^2) = \mathbb{E}(\hat{\sigma}^2 - \sigma^2)^2$$
The minimum loss over $\mathcal{M}$ is achieved by

$$\hat{\sigma}_*^2 = \frac{n - 1}{n + 1} S^2$$

But note asymmetry of loss-function w/r to under- and over-estimation. Instead, Stein’s loss

$$L_{St}(\hat{\sigma}^2, \sigma^2) = \frac{\hat{\sigma}^2}{\sigma^2} - \log(\frac{\hat{\sigma}^2}{\sigma^2}) - 1$$

has

$$\hat{\sigma}_*^2 = S^2$$

as minimizer over $\mathcal{M}$
Loss-function $d_F$

- Define loss as

$$d_F(\hat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}}) = \sqrt{(1/2) \sum \log^2(\sigma_i)},$$

where $\sigma_i$ = generalized eigenvalues of $(\hat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}})$

- Why this choice?
  - It is a metric (Förstner) for SPD matrices
  - Invariance properties: (precision, reparametrization)
    $$d_F(A, B) = d_F(A^{-1}, B^{-1}), \quad d_F(A, B) = d_F(MAM^\top, MBM^\top)$$
  - It is Rao’s geodesic distance between Gaussians with common mean
  - Minimizer also optimal for Hellinger & Kullback-Leibler
Optimality result

\[ d_F(\hat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}}) = \min_{\mathcal{M}_r} \quad \text{over } \mathcal{M}_r \text{ with} \]

\[ \hat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{r} \frac{\delta_i^2}{1 + \delta_i^2} w_i w_i^\top \]

\((\delta_i^2, w_i) = \text{generalized eigen-pair of } (H, \Gamma_{\text{pr}}^{-1})\)

with loss

\[ d_F(\hat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}}) = \sum_{i > r} \log(1 + \delta_i^2)^2 \]
Interpretation: Variance ratios

- $\tilde{w}_i = \Gamma_{pr}^{-1} w_i$ are minimizers of

\[
R(z) = z^\top \Gamma_{pos} z / z^\top \Gamma_{pr} z = \frac{\text{Var}(z^\top x | y)}{\text{Var}(z^\top x)}
\]

and maximizers of $1 - R(z)$ and:

\[
R(\tilde{w}_i) = \frac{\text{Var}(\tilde{w}_i^\top x | y)}{\text{Var}(\tilde{w}_i^\top x)}
\]

\[
1 - R(\tilde{w}_i) = \frac{\text{Var}(\tilde{w}_i^\top x) - \text{Var}(\tilde{w}_i^\top x | y)}{\text{Var}(\tilde{w}_i^\top x)}
\]

\[
\hat{\Gamma}_{pos} = \Gamma_{pr} - \sum_{i=1}^{r} \left[ 1 - R(\tilde{w}_i) \right] w_i w_i^\top
\]

\[
\hat{\Gamma}_{pos}^{-1} = \Gamma_{pr}^{-1} + \sum_{i=1}^{r} R(\tilde{w}_i) \tilde{w}_i \tilde{w}_i^\top
\]
Interpretation: Projectors

Projector (i.e., $P_r^2 = P_r$):

$$P_r = \sum_{i=1}^{r} w_i \tilde{w}_i^\top$$

Projected Bayesian model: (reduced model)

$$y \mid x \sim N(G P_r x, \Gamma_{\text{obs}}), \quad x \sim N(0, \Gamma_{\text{pr}})$$

Posterior matrix $= \hat{\Gamma}_{\text{pos}}$

Posterior mean is also an optimal low-rank approximation of $\mu_{\text{pos}}(y)$
vs optimaility in Frobenius norm

- \( \| \hat{\Gamma}_{\text{pos}} - \Gamma_{\text{pos}} \|_{\text{Frob}} = \min. \) over \( \mathcal{M}_r \) with

\[
\hat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{r} \lambda_i^2 u_i u_i^\top,
\]

\((\lambda_i^2, u_i) = \text{gen. eigen-pair of } (\Gamma_{\text{pr}}^{1/2} G^\top \Gamma_y^{-1} G \Gamma_{\text{pr}}^{1/2}, \Gamma_{\text{pr}}^{-1})\)

- \( \{u_i\} \) maximize \( V(z) = z^\top (\Gamma_{\text{pr}} - \Gamma_{\text{pos}}) z \) and

\[
V(u_i) = \text{Var}( u_i^\top x ) - \text{Var}( u_i^\top x \mid y )
\]

\[
\hat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{r} V(u_i) u_i u_i^\top
\]
vs optimality in Frobenius norm (cont.)

- **Frobenius**: directions that maximize difference:
  \[ \text{prior variance} - \text{posterior variance} \]

- ** Förstner**: directions that maximize relative difference:
  \[ \frac{\text{prior variance} - \text{posterior variance}}{\text{prior variance}} \]

- **Frobenius**: no optimality in Hellinger distance or Kullback-Leibler divergence

- **Frobenius**: more difficult eigenvalue problem
vs Hessian-based approximation

- Recall: \( \Gamma_{\text{pos}} = (H + \Gamma_{pr}^{-1})^{-1} \)

- Use low-rank approximation of \( H \) in Frobenius norm:

\[
\hat{H} = \sum_{i=1}^{r} s_i^2 v_i v_i^\top,
\]

\((s_i^2, v_i) = \text{eigen-pairs of } H\)

which defines low-rank approximation of \( \Gamma_{\text{pos}} \):

\[
\hat{\Gamma}_{\text{pos}} = (\hat{H} + \Gamma_{pr}^{-1})^{-1} = \Gamma_{pr} - KK^\top \in \mathcal{M}_r
\]

\(\Rightarrow\) Sub-optimal w/r Förstner
vs prior-based approximation

- Use principal components \((t_i^2, u_i)\) of \(\Gamma_{pr}\) to define rank-\(r\) orthogonal projector

\[
P_r = U_r U_r^T
\]

- Projected Bayesian model:

\[
y \mid x \sim N(G P_r x, \Gamma_{obs}), \quad x \sim N(0, \Gamma_{pr})
\]

\[\Rightarrow \hat{\Gamma}_{pos} \equiv \text{Posterior matrix}\]

\[
\hat{\Gamma}_{pos} = \Gamma_{pr} - KK^\top \in \mathcal{M}_r
\]

\[\Rightarrow \text{Sub-optimal w/r Förstner} \]
Better to compare $\Gamma_{\text{pos}}$ and $\hat{\Gamma}_{\text{pos}}$ accounting for both likelihood and prior.

What if the actual QoI is $Q(x)$ and not $x$? Need to account for likelihood, prior and QoI (goal-oriented).

What if forward problem $G$ is nonlinear? Posterior no longer characterized by $\mu_{\text{pos}}(y)$ and $\Gamma_{\text{pos}}(y)$.

What if $\Gamma_{\text{obs}}$ is not known precisely?
Goal-oriented inverse problem

- $y = \text{data}$
  $x = \text{high-dimensional parameter vector}$
  $G = \text{forward operator}$
  quantity of interest (QoI) $z = Qx$

- Same Bayesian model: assume $\Gamma_{pr} \succ 0$ and $\Gamma_{obs} \succ 0$,
  $y \mid x \sim N(Gx, \Gamma_{obs}), \quad x \sim N(0, \Gamma_{pr})$

  Posterior: $z \mid y \sim N( Q \mu_{\text{pos}}(y), \ Q \Gamma_{\text{pos}} Q^\top )$

- But approximation based on $x$ may not be good for $z$.  
  $\Rightarrow$ Need posterior approximations tailored to QoI
Naive approximation

- Since $\Gamma_{Z|Y} = Q\Gamma_{pos}Q^\top$, use previous low-rank approximation $\hat{\Gamma}_{pos}$ and define

$$\hat{\Gamma}_{Z|Y} = Q\hat{\Gamma}_{pos}Q^\top$$

- This approximation is not tailored to the QoI
  Important directions for $x$ may not be important for $z$ (example below)

$\Rightarrow$ Redefine the model
Inherited linear model for QoI

- Distribution of \((y, x)\) defines a distribution of \((y, z)\):
  - Prior: \(z \sim N(0, \Gamma_Z)\), \(\Gamma_Z = Q \Gamma_{pr} Q^\top\)
  - Likelihood:
    \[
    \mathbb{P}[Y \in B \mid X] = \int_B f_{Y \mid X}(y \mid X) \, d\nu(y) \quad \Rightarrow
    \]
    \[
    \mathbb{P}[Y \in B \mid Q(X)] = \int_B \mathbb{E} \left[ f_{Y \mid X}(y \mid X) \mid Q(X) \right] \, d\nu(y).
    \]
- Easy for the Gaussian linear model:
Likelihood, approximation class for QoI

\[ y \mid z \sim N(G_Q \Gamma_z^{-1/2} z, \Gamma_\Delta) \]

\[
G_Q = G \Gamma_{pr} Q^\top \Gamma_z^{-1/2}
\]

\[
\Gamma_\Delta = \Gamma_{obs} + G (\Gamma_{pr} - \Gamma_{pr} Q^\top \Gamma_z^{-1} Q \Gamma_{pr}) G^\top
\]

\[
= \Gamma_y - G_Q G_Q^\top
\]

- Use previous optimal low-rank approximations with new likelihood and prior

- Approximation class

\[
\mathcal{M}^Z_r = \{ \Gamma = \Gamma_z - KK^\top : \Gamma \succ 0, \text{ rank}(K) \leq r \}
\]
Optimality result

\[ d_F(\hat{\Gamma}_{Z|Y}, \Gamma_{Z|Y}) = \min_{\mathcal{M}_r^Z} \text{ over } \mathcal{M}_r^Z \text{ with } \]

\[ \hat{\Gamma}_{Z|Y} = \Gamma_Z - \sum_{i=1}^{r} \lambda_i \hat{q}_i \hat{q}_i^\top, \quad \hat{q}_i = G_Q \Gamma_Z^{1/2} q_i \]

\((\lambda_i^2, q_i) = \text{generalized eigen-pair of } (G_Q G_Q^\top, \Gamma_y)\)

with loss

\[ d_F(\hat{\Gamma}_{Z|Y}, \Gamma_{Z|Y}) = \frac{1}{2} \sum_{i>r} \log(1 - \lambda_i^2)^2 \]
Example

- Heat transfer problem. Infer temperature of a CPU ($\mathcal{D}_1$) from noisy temperatures in heat-sink ($\mathcal{D}_3$) connected to CPU by a silicon layer ($\mathcal{D}_2$)

$x = \text{initial temp. on } \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$

Qol: $z = x$ on $\mathcal{D}_1$
Extensions

- Applications to OED (ongoing)
- Applications to stochastic approximations (ongoing)
- Nonlinear case (ongoing)
- $\Gamma_{\text{obs}}$ is often unknown $\Rightarrow$ model parametrically as $\Gamma_{\text{obs}}(\theta)$ with unknown $\theta$

(2) Hierarchical Bayes: model $\theta$ as random
Posterior sampling:

$$f(x | \theta, y) \propto f(y | x, \theta) f(x)$$
$$f(\theta | x, y) \propto f(y | x, \theta) f(\theta)$$

(3) Empirical Bayes: use estimate $\hat{\theta}$ as fixed
References


